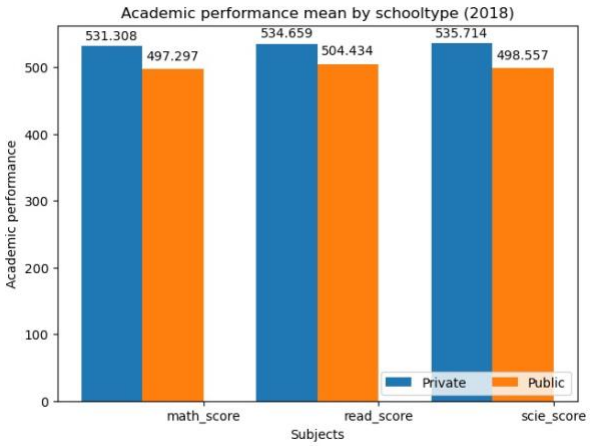
There are various factors that contribute to student's learning. To gain a clear insight on the determinants of student's learning, a data set derived from the OECD's Program for International Student Assessment (PISA) will be analysed. Through a process of conducting data analysis both numerically and visually will be essential to uncover the underlying influences. More importantly to answer the research question, "Do private school pupils outperform public school pupils in Math, Reading or Science tests?"

The main variables of interest from the PISA data set to first address the research question, are School Ownership (*schltype*) and student's test scores: Mathematics tests(*math_score*), Reading tests (*read score*), Science tests(*scie_score*). Beginning the data analysis, the data set has been grouped based on their school ownerships, public and private school. To ensure that comparison data is equal between the two school types, private school data has been limited to 2304 observations. The main summary statistic highlighted is the average test scores, which has been calculated and visualised in the bar chart below.



It is evident that the private school pupils do outperform public school pupils across the three subjects. Though this builds on the hypothesis that academic performance between private and public-school pupils are not equal, is the variation statistically significant?

A statistical T-test is carried out to analyse this hypothesis and to determine to what extent does private school academic performance differ from that of public pupils. The hypothesis are broken down as the null hypothesis(there is no difference in the mean academic performance) and alternative hypothesis (there is a difference in the mean academic performance). The T-test is carried out and compared for each subject individually. To summarise the results, the t-test across all subjects (math, reading and science) range from 11.2797 to 14.6884 which is large in absolute value, suggesting that there is substantial difference between private school pupils' academic performance in comparison to public schools. Whilst the p-value associated ranges from 1.5e-47 to 9.2e-48 which is extremely small compared to the significance level $\alpha=0.05$. In that case, the is evidence to reject the null hypothesis and accept the alternative hypothesis as true.

It is keen to consider that there are other variables that may contribute to this difference in academic performance between students. These variables can be categorized within the home background of the students. For example, the variable, home educational resources(*hedres*) can have a significant impact on academic performance. This is because, the presence of a study desk, a quiet place, and books to help with schoolwork, can help pupils practice further what they have been taught at school. Further contributing to their likelihood in performing better and improving in their test scores. As a result, school ownership should not be the only independent variable that is considered to contribute to academic performance.

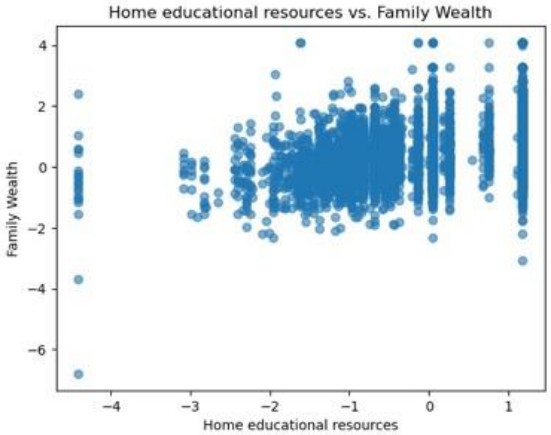
A linear regression model will be vital to understand fully the scope of the relationship between the dependent variable (academic performance) and independent variables (school ownership and home educational resources). The individual test scores in math, reading and science have been combined into a new variable *total_score* and will act as the dependent variable. The linear regression model to describe the relationship between total score, school ownership and home educational resources is:

$\text{Test scores}_i = \beta_0 + \beta_1 \text{schltype}_i + \beta_2 \text{hedres}_i + \epsilon_i$. The estimated parameters have been calculated and displayed below.

```
Intercept                          1556.202161
schltype[T.Private Independent]     166.725270
schltype[T.Public]                  -56.216277
hedres                               44.365141
dtype: float64
```

From the values above, the variable *hedres*, has a positive relationship with *total_score*. Whereby with increased home educational resources leads to a positive increase in total score of students of 44.365. Furthermore, it can be concluded that attending a private independent school increases the total score of a student by 166.725 while attending a public school decreases the total score of a student by 56.216. This emphasises that there is a positive relationship between total score and private school and a negative relationship between test score and public school.

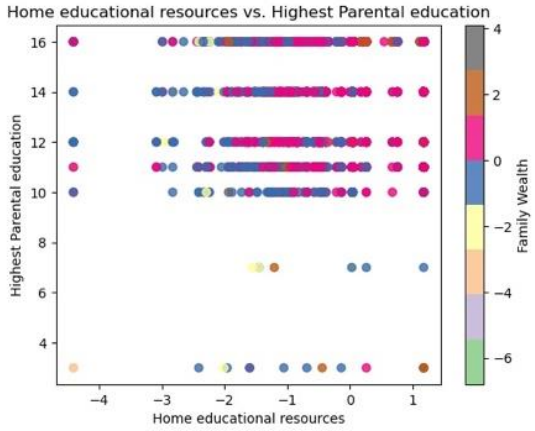
Taking a further step to consider whether there is potential multicollinearity between hedres and other home background variable indicators. It is fair to assume that with increased family wealth could potentially increase the access to home educational resources such as books, technology equipment and quiet study area. This is because there is likely more disposable income to invest in resources. Therefore, we can consider family wealth*(wealth*) as a potential omitted variable. To investigate this, the correlation between *hedres* and *wealth* has been calculated to be 0.344. Which suggests that there is a positive relationship but with moderate magnitude. It is visually represented in the scatter graph below.



As displayed, majority of the values in the scatter graph are dispersed on the top half of the scatter graph. Which suggests that pupils from a higher socioeconomic background have access to home educational resources. Concluding that there is a positive linear relationship between hedres and wealth that should be considered.

Additionally, the variable *pared,* an index that indicates the highest educational background for parent, could have a great effect in the home educational resources. As it is possible that, a more educated parental figure understands the importance of fostering a good study environment from their own academic experience and would invest in books, internet etc.

Furthermore, the more educated a parent is the more likely they are to be qualified for a higher paying job and therefore increasing the family wealth. To understand this relationship, a scatter graph is created of home educational resources vs highest parental education, including family wealth as additional variable of interest.



It is evident that there are higher values of access to home educational resources for families with highest parental education being 10 years and more. Among the same demographic, their family wealth is also majority positive. Indicating that the higher the parental educational background the more likely the family wealth index and as a result a higher *hedres* index.

From understanding the relationship between *hedres* and other variables *wealth* and *pared* it is interesting to see if there is a correlation with the main variable, *schltype*. Below is a table of calculated correlation between school type and additional variables (*hedres, wealth, pared)*. The coefficients are possible however they are also small in absolute values.

		hedres	wealth	pared
schltype				
Private Government-dependent	**hedres**	1.000000	0.324951	0.258758
	wealth	0.324951	1.000000	0.217173
	pared	0.258758	0.217173	1.000000
Private Independent	**hedres**	1.000000	0.251470	0.087131
	wealth	0.251470	1.000000	0.043162
	pared	0.087131	0.043162	1.000000
Public	**hedres**	1.000000	0.348505	0.225340
	wealth	0.348505	1.000000	0.236994
	pared	0.225340	0.236994	1.000000

The original linear regression model is modified to include the additional variables. This new linear regression model is expressed as:

$$\text{Test scores}_i=\beta_0+\beta_1\text{schltype}_i+\beta_2\text{hedres}_i+\beta_3\text{wealth}_i+\beta_4\text{pared}_i+\epsilon_i.$$

The new estimated coefficients are shown below

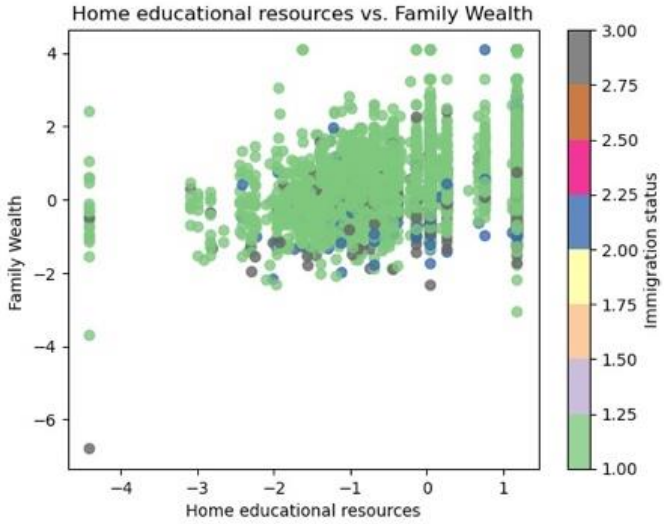
```
Intercept                                1373.009116
C(schltype)[T.Private Independent]        156.390644
C(schltype)[T.Public]                     -56.577193
hedres                                     39.762719
wealth                                     -6.365919
pared                                      13.311845
dtype: float64
```

In comparison to the original linear regression model, the *intercept*, *private independent* and *hedres* estimation coefficients have decreased. This suggests that their influence on the

students' *total_score* has decreased. It also keen to note that, the estimation coefficient has remained unchanged, which suggests that the additional variables have no effect on students' academic performance in public schools. Which confirms that there was omitted variable bias, that led to an overestimation of the effects of private school and home educational resources on total scores. Concluding that it is vital to include *wealth* and *pared* as additional variables of interest.

Another perspective of considering the parents influence on the students' academic performance is their immigration status. The variable *immig* captures the immigration status of both parents and students. Whether they are a "Native", "Second-Generation" or "First-Generation". An example of the potential influence of *immig* is, students who are non-native pupils may have to adapt to a new culture, new environment as well as new language, which could reduce the amount of focus and time dedicated to studying. In terms of their parents, they may not understand the resources needed for their children to succeed in a new education system. To effectively analyse the influence of immig on total score as well as other independent variables, the *immig* variable is transformed into a numerical index from categorical. "Native", "Second-Generation" and "First-Generation" replaced by 1, 2 and 3 respectively.

From calculating the correlation coefficient between immigration and wealth and hedres. There is a negative relationship between wealth and immigration of 0.168, however not of a strong magnitude. Whilst there is a positive relationship between *hedres* and *immig* of 0.0227, almost negligible. This relationship is visualised below, a scatter plot of *hedres vs wealth* including *immig*.



It is clear, from the green colour, that a large proportion of "Native" families are from a higher socioeconomic background and therefore have access to home educational resources. As a result, it would be interesting to analyse if *immig* has a potential influence on *total_score*.

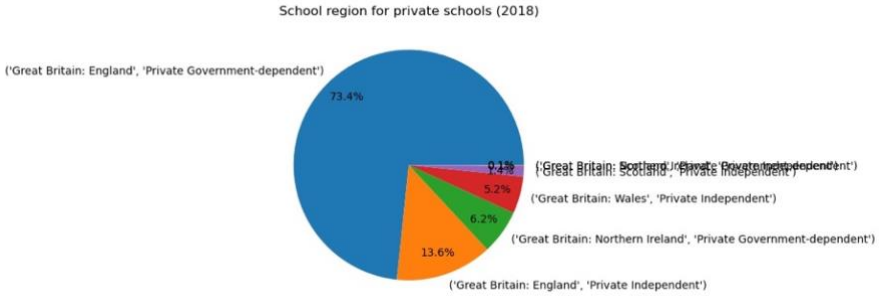
A modified linear model can be constructed, including all the discussed independent variables, expressed as:

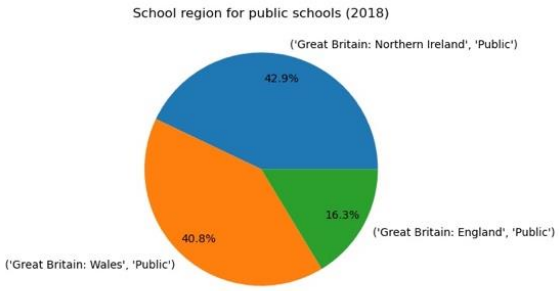
$\text{Test scores}_i = \beta_0 + \beta_1 \text{schltype}_i + \beta_2 \text{hedres}_i + \beta_3 \text{wealth}_i + \beta_4 \text{pared}_i + \beta_5 \text{immig}_i + \epsilon_i$. With The estimated coefficients as follows:

```
Intercept                              1423.597707
C(schltype)[T.Private Independent]      164.142065
C(schltype)[T.Public]                   -59.310537
hedres                                   41.554252
wealth                                  -11.957672
pared                                    13.436456
immig                                   -40.621372
dtype: float64
```

The addition of *immig* to the variable has led to an increase in the intercept and negative coefficient of wealth. This suggests that immig and wealth are in fact negatively related which leads to the presence of multicollinearity in the model. In that case, it would not be effective to include *immig* as it strongly influences existing variables which could lead to skewed results.

A limitation of this analysis is that it might suffer from potential omitted variable bias. For example, the student's revision time has not been considered which can be very influential. This is because, students might be gifted and with great potential however have not invested enough time to fully grasp the content which might affect their grades, irrespective of their school association. Moreover, the data set's total observations are from one particular year (2018). It would be beneficial to compare the PISA data across multiple years to interpret whether there is a consistent trend.
Additionally, as illustrated in the pie charts below, the region the public schools and private schools are located are not the same. This is a limitation as with different regions comes with variation in culture and regional trends that is not accounted for.





As illustrated above, majority of public schools are in Northern Ireland whilst most private schools are in England, as a result, it is likely that there are external factors that would affect one region and not the other. However, within this data set, the difference between the average academic performance of Northern Ireland and England is 57.824. It can be considered as negligible and therefore, school region does not have a strong effect on academic performance.

In conclusion from the data analysis, private school pupils do outperform public school pupils across the three subjects Math, Reading and Science tests. Showcased by the differences in the mean test scores. However, it would not be accurate to suggest that the school ownership is the main reason for the disparities in academic performance. Home background factors such as presence of home educational resources, highest parental education and family wealth are a

determining factors of pupil's academic performance. Despite the influence of other variables in determining student's performance, the type of school association remains to have the greatest influence.